Lecture 12 - EM & GMM

Friday, October 9, 2020 13:28

today : - EM algorithm
- GMM

EM – maximum likelihood in latent variable model

Setup :

X observed variable

Z latent variable

log-likelihood $\log p(x_{1:n}; \theta) = \log \left(\prod_{i=1}^{n} p(x_i; \theta) \right)$

$= \sum_{i=1}^{n} \log p(x_i; \theta)$

$= \sum_{i=1}^{n} \log \left[\sum_{z} p(x_i, z_i; \theta) \right]$

problem? → gives multi-modal opt. problem (non-convex)

options for ML in latent variable model

1) do gradient ascent on a non-concave obj.

f is convex $\Leftrightarrow$ $-f$ is concave

2) EM alg. → block-coordinate ascent on auxiliary fct. which lower bounds $\log p(x_{1:n}; \theta)$

nice interpretation in terms of filling "missing data"

ie. E step → fill z with "soft-values"

M step → max w.r. to θ for fully observed model

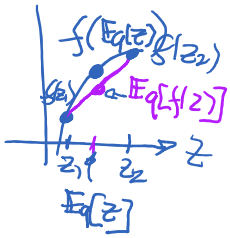
trick overview:

$\log \sum_z p(x,z) = \log \sum_z q(z) \frac{p(x,z)}{q(z)}$

$= \log \mathbb{E}_q \left[\frac{p(x,z)}{q(z)} \right]$

Jensen's inequality

$\mathbb{E}_q[f(\frac{g(z)}{z})] \leq f(\mathbb{E}_q[\frac{g(z)}{z}])$ when f is concave



Jensen's inequality trick

$\mathbb{E}_q \left[\log \frac{p(x,z)}{q(z)} \right] = \sum_z q(z) \log p_\theta(x,z) - \sum_z q(z) \log q(z)$

$\triangleq \mathcal{L}(q, \theta) \triangleq \mathbb{E}_q [\log p(x, z; \theta)] + H(q)$

"expected complete log-likelihood"

"entropy of q"

"expected complete log-likelihood" "entropy of q"

we have $\log p(x;\theta) \geq \mathcal{L}(q,\theta) \quad \forall q;\theta$

EM algorithm: E step: $q_{t+1} \triangleq \underset{q \text{ (dist. over } z)}{\operatorname{argmax}} \mathcal{L}(q,\theta_t) \Rightarrow q_{t+1}(z) = p(z|x;\theta_t)$

M step: $\theta_{t+1} \triangleq \underset{\theta}{\operatorname{argmax}} \mathcal{L}(q_{t+1},\theta)$

$= \underset{\theta}{\operatorname{argmax}} \mathbb{E}_{q_{t+1}(z)}[\log p(x,z;\theta)]$

this is another ML problem, but for complete information

(often, replace z with $\mathbb{E}_q[z]$ in this expression)

from Jensen's ineq.

* we had $\log p(x;\theta) \geq \mathcal{L}(q,\theta)$ $\quad \log(\mathbb{E}_q[g(z)]) \geq \mathbb{E}_q \log(g(z))$

in Jensen's ineq., you get strict ineq. unless the dist. $g(z)$ is degenerate i.e. takes only one value (when f is strictly concave)

i.e. when $g(z)$ = constant

$g(z)$ = constant above i.e. $\frac{p(x,z)}{q(z)} = \text{const.} \ \forall z \Rightarrow q(z) \propto p(x,z)$

i.e. $\boxed{q(z) = p(z|x;\theta)}$ makes equality $\mathcal{L}(q,\theta) = \log p(x;\theta)$

$\mathcal{L}(q_{t+1},\theta_t) = \log p(x;\theta_t) \geq \mathcal{L}(q,\theta_t)$ i.e. $\underset{q}{\operatorname{argmax}} \mathcal{L}(q,\theta_t) = p(z|x;\theta_t)$

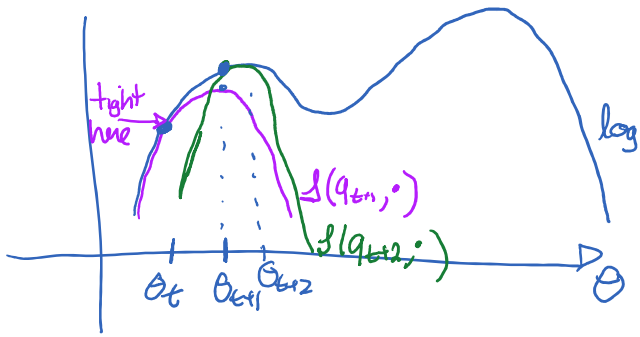
$\Rightarrow$ q_{t+1} maximizes $\mathcal{L}(q,\theta_t)$ w.r. to q and $\mathcal{L}(q_{t+1},\theta_t) = \log p(x;\theta_t)$

properties:

a) $\log p(x;\theta_{t+1}) \geq \log p(x;\theta_t)$ proof: $\log p(x;\theta_{t+1}) \geq \mathcal{L}(q_{t+1},\theta_{t+1})$

$\geq \mathcal{L}(q_{t+1},\theta_t)$

$= \log p(x;\theta_t)$



b) θ_t in EM converges to a stationary pt of $\log p(x;\theta)$

ie. $\nabla_\theta \log p(x;\theta)\big|_{\hat{\theta}} = 0$

like K-means, initialization is crucial
→ usually do random restarts

for GMM,
could use k-means++ to initialize the μ_k's

14h34

c) $\mathcal{L}(q,\theta) = \mathbb{E}_q\left[\log \frac{p(x,z;\theta)}{q(z)}\right]$

$p(z|x;\theta)$

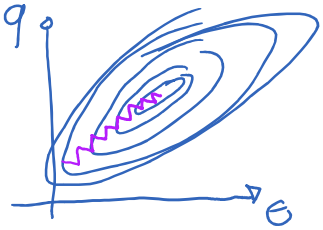
$$\log p(x;\theta) - \mathcal{L}(q,\theta) = -\mathbb{E}_q\left[\log \frac{p(z|x;\theta)}{q(z)} \frac{p(x;\theta)}{p(x;\theta)}\right]$$

$$= \mathbb{E}_q\left[\log \frac{q(z)}{p(z|x;\theta)}\right] \triangleq KL(q(\cdot) \| p(\cdot|x;\theta))$$

$\log p(x;\theta) - \mathcal{L}(q,\theta) = KL(q \| p(\cdot|x,\theta))$

KL divergence

we will revisit for variational inference $q \in Q$



for GMM model

$z_i \sim Mult(\pi)$

$x_i | z_{ij}=1 \sim N(\mu_j, \Sigma_j)$

shorthand to say $z_{ij}=1$

$\theta = (\pi, (\mu_j)_{j=1}^K, (\Sigma_j)_{j=1}^K)$

notation: $x = x_{1:n}$

$z = z_{1:n}$

exercise:

$$p(z|x) = \prod_i p(z_i|x) = \prod_i p(z_i|x_i)$$

complete log-likelihood:

$$\log p(x,z;\theta) = \sum_i \left[\log p(x_i|z_i;\theta) + \log p(z_i;\theta)\right]$$

$$\log p(x_i, z_i; \theta) = \sum_{i=1}^{n} \left[\log p(x_i | z_i; \theta) + \log p(z_i; \theta) \right]$$

($\log p(x_i | z_i; \theta)$: Gaussian; $\log p(z_i; \theta)$: multinoulli)

$$= \sum_{i=1}^{n} \left[\sum_{j=1}^{K} z_{ij} \log N(x_i | \mu_j, \Sigma_j) + \sum_{j=1}^{K} z_{ij} \log \pi_j \right]$$

$$\mathbb{E}_q [\log p(x, z; \theta)] = \sum_{i=1}^{n} \sum_{j=1}^{K} \mathbb{E}_q[z_{ij}] \left(\log N(x_i | \mu_j, \Sigma_j) + \log \pi_j \right)$$

$\mathbb{E}_q[z_{ij}] = q(z_{ij} = 1)$ [marginal distribution]

during EM, $q_{t+1}(z) = p(z | x; \theta_t)$

weight $r_{ij}^t \triangleq p(z_{ij} = 1 | x_i; \theta_t) = q_{t+1}(z_{ij} = 1)$

E-step is computing $q_{t+1}(z) \triangleq p(z | x; \theta_t)$

$$= \prod_i p(z_i | x_i; \theta_t)$$

$$\Rightarrow q_{t+1}(z_i) \propto p(x_i | z_i; \theta_t) \, p(z_i; \theta_t)$$

($p(x_i | z_i; \theta_t)$: Gaussian; $p(z_i; \theta_t)$: $\pi_{z_i}^t$)

$$r_{ij}^t = q_{t+1}(z_{ij} = 1) = \frac{\pi_j^{(t)} N(x_i | \mu_j^{(t)}, \Sigma_j^{(t)})}{\sum_{l=1}^{K} \pi_l^{(t)} N(x_i | \mu_l^{(t)}, \Sigma_l^{(t)})}$$

numerator: $p(x_i, z_{ij} = 1 | \theta^{(t)})$; denominator: $p(x_i | \theta^{(t)})$

E step for GMM: compute $r_{ij}^{(t)}$ for $i = 1, \dots, n$ using $\theta^{(t)}$

M step: $\max_{\{\mu_j, \Sigma_j, \pi_j\}} \sum_i \sum_j r_{ij}^{(t)} \left[\log p(x_i | \mu_j, \Sigma_j) + \log \pi_j \right]$

exercise:

M step for EM for GMM

$$\hat{\pi}_j^{(t+1)} = \frac{\sum_i r_{ij}^{(t)}}{n}$$ ("soft-count")

$$\hat{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^{n} r_{ij}^{(t)} x_i}{\sum_{i=1}^{n} r_{ij}^{(t)}}$$

$$\hat{\Sigma}_j^{(t+1)} = \frac{\sum_{i=1}^{n} r_{ij}^{(t)} (x_i - \hat{\mu}_j^{(t+1)})(x_i - \hat{\mu}_j^{(t+1)})^T}{\sum_{i=1}^{n} r_{ij}^{(t)}}$$

$$\Sigma_j^{(t+1)} = \frac{\sum_{i=1}^{n} r_{ij}^{(t)} (x_i - \mu_j^{(t+1)})(x_i - \mu_j^{(t+1)})^\top}{\sum_{i=1}^{n} r_{ij}^{(t)}}$$

initialize: e.g. $\mu_j^{(0)}$ from k-means++

$\Sigma_j^{(0)}$ big spherical covariance $\Sigma_j^{(0)} = \sigma^2_{\text{big}} I$

$\pi_j^{(0)}$: proportions from k-means++

EM step in GMM with fixed $\Sigma_j = \sigma^2 I$ with $\sigma^2 \to 0$

$\rightsquigarrow$ get k-means alg?